Use of Scanner Data in Measuring the Consumer Price Index in the Conditions of the Slovak Republic

Helena Glaser - Opitzová[1]

[1] University of Economic, Faculty of Economic Informatics/Department of Statistics,
Dolnozemská cesta 1, Bratislava, 852 32
Slovak Republic

Helena.Glaser-Opitzova@statistics.sk

https://doi.org/10.53465/EDAMBA.2021.9788022549301.92-102

Abstract. The compilation of the Consumer Price Index (CPI) is, in the conditions of the Slovak Republic, based on the fixed consumer basket. Only the lockdown during pandemics showed us how a fixed basket can very quickly become irrelevant due to a rapidly changing model of consumer behaviour. The changeover from traditional data collection to the usage of data from scanners practically means the passing from static universe of selected sorts of goods to the dynamic universe of all goods consumed. While classical bilateral indices may be appropriate for a fixed basket, the transition to a dynamic approach raises the question of whether traditionally used methods are still valid. Our goal is to publish high quality index and therefore it is necessary to bear in mind that the index number formula selection and methods applied to data from scanners can have a significant impact on achieved results. The paper presents the basic conceptual framework for the use of scanner data for the purposes of compiling the CPI in relation to selected findings of the experimental study performed on real data of five retail chains for the "Food and non-alcoholic beverages" division.

Keywords: Scanner Data, Consumer Price Index, Index Number Formula

JEL classification: *E 43, E 31*

1 Introduction

The Consumer Price Index (CPI) is considered to be one of the most important socio-economic indicators produced in official statistics. The CPI expresses the average change in prices of goods and services purchased by population for consumption.

In the conditions of the Slovak Republic, the data needed for CPI compilation, have been collected through field data collection. During the first twenty days of each month, the employees of the Statistical Office of the Slovak Republic visit shops and service establishments and record so-called counter prices, prices for which are goods and services offered to consumers.

Within the field collection, those shops and service establishments, where people usually go for shopping, are visited. The development of prices is monitored via fixed consumer basket of goods and services (728 price representatives) of constant quality and similar characteristics, which are selected to represent household expenditure over a year. In this above-mentioned set of representatives, there are 146 items that belong to the area of food and non-alcoholic beverages, which is, in the conditions of the Slovak Republic, the second most expensive area of household consumption following housing costs. At present, this traditional method is gradually replaced in the world, especially in the area of food and non-alcoholic beverages by data from scanners, which are defined by [7] as "*detailed data on sales of consumer goods obtained by scanning the bar codes for individual products at electronic points of sale in retail outlets*."

This type of transaction data offers more precise and detailed information, and unlike the previous method of obtaining data it also contains quantities of individual goods that are sold in actual period.

The Statistical Office of the Slovak Republic gradually concluded the agreements on the mutual cooperation in the field of statistics with five largest retail chains that have provided the SOSR with the access to this type of data since 2018.

Even though this data is not currently implemented into routine production of CPI, at the pandemic time it represented an important supplementary or substitute source of information because data on prices of some goods could not be obtained without potential threat to employees of the SOSR. It can also be assumed that the consumer behaviour of the population of the Slovak Republic was being changed during the lockdown, similarly to the consumer behaviour of the population of other European countries [9].

Changeover from the traditional collection of data on prices of goods and services to the usage of data from scanners is demanding and long-term process that involves various activities, from negotiations with representatives of retail chains on data provision to addressing methodological and practical issues and testing of different calculation methods in the production environment. The article demonstrates the basic concept of possible use of data from scanners in the field of official price statistics and the first findings of the experimental study performed on real data of five retail chains for food and non-alcoholic beverages in the conditions of the Slovak Republic.

2 Conceptual framework

2.1 Scanner data advantages and disadvantages

Scanner data from a specific vendor and for a given time period represents an exhaustive list of all item codes, their turnovers, and quantities sold. They enable to compile an index from all vendor´s transactions or transactions of the store. For example, the assortment of food and non-alcoholic beverages, which is currently the subject of an empirical study at the Statistical Office of the Slovak Republic, is covered, in the conditions of Slovakia, by 7,000 to 29,000 items, depending on the retail chain.

Scanner data enables to statisticians to use what has been sold actually and to include many more items in the CPI calculation in comparison with traditional price collection. It also means that if the information on turnovers is available, weights can be assigned to individual items. The undeniable advantage of scanner data is also a large amount of detailed information about individual products[1], which enables to define the so-called homogeneous groups of products and calculate the price indices of elementary aggregates at a lower level of aggregation than COICOP5[2].

Implementing scanner data in the production of price statistics can also save the cost of traditional price data collection.

Scanner data" reflects the dynamics of actual purchases in each elementary aggregate, as each transaction is recorded. The appearance of new item codes, the disappearance of item codes and changes of their relative importance are visible in the data set. It follows from the above-mentioned, that the scanner data, compared to the traditional data source, has both advantages and disadvantages for the production of the CPI. There is some additional work burden for the NSIs due to increased need for data cleaning and processing, as well as initial cost of setting up the IT system. The dependence of the NSI on individual retail chains is increasing, which means that non-delivery of data will have greater consequences than before. More advanced IT competencies are required from the CPI staff

2.2 Items selection

In [4], two methods of selecting items that enter the calculation of price indices at the level of elementary aggregates are generally recommended, namely the static approach and the dynamic approach. The static approach simulates the traditional fixed consumer basket, with the difference that the prices obtained by the traditional survey are replaced by average prices per unit of goods. Information on actual sale is utilized only at initial selection of items into consumption basket or their replacement in the course of a year.

In dynamic method, the index calculation includes a representative selection of items for every two consecutive months after filtering items out with extreme price changes, items of goods on clearance sales and low sales goods. It means, that the index at the level of elementary aggregate is calculated on the basis of a set of matched representative item codes for items that are actually sold in two consecutive periods.

2.3 Elementary price indices

An elementary price index is the price index calculated for the elementary aggregate. An elementary aggregate consists of a relatively homogeneous set of goods or services with similar expected price changes. They can cover the whole country, individual regions, or it is possible to define them for different types of outlets. Various different methods and formulae may be used to calculate elementary price indices. The choice of index number formula and calculation method can have a significant effect on the results obtained. The European Commission recommends the Jevons price index at the

[1] This information applies if the retail chain has a quality internal classification and is willing to share it.

[2] COICOP The Classification of individual consumption by purpose

lowest level of data aggregation [3] (see also [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]), which can be written for the base period (0) and current period (t) as follows:

$$P_J^{0,t} = \left(\prod_{i \in S} \frac{p_i^t}{p_i^0}\right)^{1/N_{0,t}} = \frac{(\prod_{i \in S} p_i^t)^{1/N_{0,t}}}{(\prod_{i \in S} p_i^0)^{1/N_{0,t}}} \tag{1}$$

S represents the set of identical items of goods belonging to a certain category and $N_{0,t}$ the number of identical items of goods. p_i^0 a p_i^t are prices (prices per a unit of goods) of each item $i \in S$ in the period *0* a *t*.

Theoretically, for homogeneous groups of products, it is also possible to use the Dutot index, which can be expressed as follows [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]:

$$P_D^{0,t} = \frac{\sum_{i \in S} p_i^t}{\sum_{i \in S} p_i^0} \tag{2}$$

The authors of some studies recommend, in some exceptional cases, to consider the use of the Carli index [**Chyba! Nenašiel sa žiaden zdroj odkazov.**], which is defined as a simple or unweighted arithmetic average of relative prices or price ratios for the period 0 a t, which can be expressed:

$$P_C^{0,t} = \frac{1}{N_{0,t}} \sum_{i \in S} \frac{p_i^t}{p_i^0} \tag{3}$$

The paper [**Chyba! Nenašiel sa žiaden zdroj odkazov.**] presents many studies proving significant differences in results between Dutot, Carli and Jevons indices.

When selecting an appropriate index for measuring CPI / HICP, two main approaches can be used, axiomatic and economic approach. From the axiomatic point of view, Jevon index is the index with the best properties. Although it has not been widely used until recently, the trend of its use by statistical offices is on the rise.

The economic approach is based on the economic theory of the consumer behaviour. The economic approach assumes that quantities consumed are a function of the prices, and data recorded arises as solutions of different problems of economic optimization of the consumer. According to this approach, the CPI is defined as a cost-of-living index (COLI). A growing number of economists and other users prefer superlative indices for CPI compilation purposes, such as e.g. Fisher index, because they consider them to be the best COLI approximation.

2.4 Weighted price index formulas

The CPI is currently still determined by means of Laspeyres index. The Laspeyres index, which uses only weights from the basic period, does not reflect changes in consumer behaviour during the reference period and it may be distorted due to substitution of goods. The Laspeyres price index [**Chyba! Nenašiel sa žiaden zdroj odkazov.**] can be expressed as follows:

$$P_{La}^{0,t} = \frac{\sum_{i \in S} p_i^t q_i^0}{\sum_{i \in S} p_i^0 q_i^0} \tag{4}$$

where p_i^t is the price of *i- th* product at time *t.*

The Paasche price index [**Chyba! Nenašiel sa žiaden zdroj odkazov.**] uses the quantity of goods sold from current period and can be written as follows:

$$P_{Pa}^{0,t} = \frac{\sum_{i \in S} p_i^t q_i^t}{\sum_{i \in S} p_i^0 q_i^t} \tag{5}$$

The above-mentioned economic approach in the price index theory assumes that the real value of the COLI belongs to the interval whose upper and lower limit is determined by the values of the Paasche price index and Laspeyres price index.

As it was mentioned in the previous subchapter, the most preferred indices for the purpose of measuring CPI are superlative price indices, firstly proposed by [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]. They can be written as follows:
The Fischer price index [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]

$$P_F^{0,t} = \sqrt{P_{La}^{0,t}\, P_{Pa}^{0,t}} \tag{6}$$

The Törnqvist price index [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]

$$P_T^{0,t} = \left(\prod\nolimits_{i \in S} \frac{p_i^t}{p_i^0}\right)^{s_i^0 + s_i^t \big/ 2} \tag{7}$$

where $s_i^0 = p_i^0 q_i^0 / \sum_{i \in S} p_i^0 q_i^0$ and $s_i^t = p_i^t q_i^t / \sum_{i \in S} p_i^t q_i^t$ refer to the proportions of expenditure in period 0 and t; q_i^0 and q_i^t are quantities sold.

2.5 Chained indices

A chained index provides a measure of cumulated effect of successive price steps. Chain indices are the indices with moving base, id est, with the base from the values of the previous period. Any index number formula can be used for the individual links in a chained index. For example, an unweighted chained Jevons index can be written as follows [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]:

$$CP_J^{0t,mt} = P_J^{0t,1t} . P_J^{1t,2t} . \dots . P_J^{(m-1)t,mt} = \tag{8}$$

If in practice, HICP/CPI are compiled from “scanner data” it is generally recommended to use chained superlative indices due to the higher degree of matching of individual item codes between two consecutive periods and the assumption of smaller differences in price and quantity. However, this assumption does not consider the existence of clearance sales and discounts, which can significantly increase the quantity of goods sold, up to several times. When the discount period ends, the price returns to its original value. However, the situation may turn out to be that population is supplied and it will take longer time until the quantity of sold goods returns to the former/original value. Under these conditions, chained superlative indices tend to decline and only gradually return to their original level (downward chain drift) in comparison with base indices.

In [**Chyba! Nenašiel sa žiaden zdroj odkazov.**] have been proposed an approach that provides drift free, superlative-type indexes through adapting multilateral index number theory.

This procedure maximizes the number of matched items in data without the risk of introducing drift to chained time series.

2.6 Multilateral price indices

In multilateral price index methods, the aggregate price change between two comparison periods is obtained from prices and quantities observed in multiple periods including the two comparison periods. They were developed for price comparisons across countries. The best known are the GEKS method, Geary-Khamis method [**Chyba! Nenašiel sa žiaden zdroj odkazov.**], and Country-Product Dummy method.

Multilateral spatial comparison of prices can be simply adapted for comparison over time. Multilateral indices meet the transitivity requirement. The index is transitive if the index that compares periods „*a*“ a „*b*“ through period „*c*“ is identical with the index that compares periods „*a*“ a „*b*“ directly.

The idea of adapting the method to the time series context was developed by Ivancic, Diewert and Fox [**Chyba! Nenašiel sa žiaden zdroj odkazov.**].

3 Empirical study

The received raw scanner data needs to be pre-processed and classified at the level of individual items. Based on “detailed” descriptions of individual products (items) at the EAN code level or on the internal classification of retail chains products, we have defined 354 homogenous product groups for food and non-alcoholic beverages and, in such way we have created, a national, 6-digit level of ECOICOP classification for this area of consumption. ECOICOP 6-digit is common for all retail chains that currently cooperate with the SOSR, id est, the elementary aggregate can be defined both at the level of the Slovak Republic and at the level of the retail chain, but not at the level of regions.

Each product, that potentially enters the calculation of the elementary index, is subject to the assessment of its “importance”. The selection of the products that enter the calculation depends on the calculation method used and the filters applied to the data. Since in the conditions of the Slovak Republic the CPI index is compiled on a monthly basis, data filtering is performed on a set of data that would potentially enter in the calculation of the month-on-month index.

Items whose prices have been increased or decreased disproportionately in comparison with the previous period have been excluded. The limits for the identification of outliers have been set to 0.3 and 3 (outlier filter) after a thoroughgoing analysis of the annual data. Furthermore, those items, whose prices and turnovers fall significantly compared to the previous period are excluded from the calculation of the index at the elementary level. These are mainly products intended for clearance sale, if price change is ≤ 0.8 and, at the same time, the change of turnovers is ≤ 0.2. This filter is known as a dumping filter. In the case of the dynamic approach method (dynamic consumer basket, which is always created up to date for two consecutive periods), subsequently products with low sales are excluded as well on the basis of formula [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]:

$$\frac{s_i^{t-1}+s_i^t}{2} > \frac{1}{n\lambda} \qquad (9)$$

It means that the product will be included into the sample for the index calculation if the ratio of the item *i* in expenditure in months *t* a *t-1* exceeds threshold value $\frac{1}{n\lambda}$, where:

n - is the number of products,
λ - is a fixed parameter and based on empirical research λ =1.25.

The effect of data filtering can be seen in Table 1.

Data filtering impact on values of chained unweighted Jevons index can be seen in Fig.1. From Fig.1 and also from the results of the experimental study in [**Chyba! Nenašiel sa žiaden zdroj odkazov.**], it is clear, that Jevons index is sensitive to filter selection. In the specific example given, the Jevons index acquires the lowest values when it is being compiled above an unfiltered database of products, probably, because it is dominated by products with smaller monthly price changes than the average monthly price change.

Table 2. Efect of data filtering (COICOP 01, average month 2019)

Type of filter	Number of items after filtering
Without filtering	60 239
outlier filter	60 070
(outlier + dumping) filter	59 638
(outlier + dumping +low sale) filter	17 057

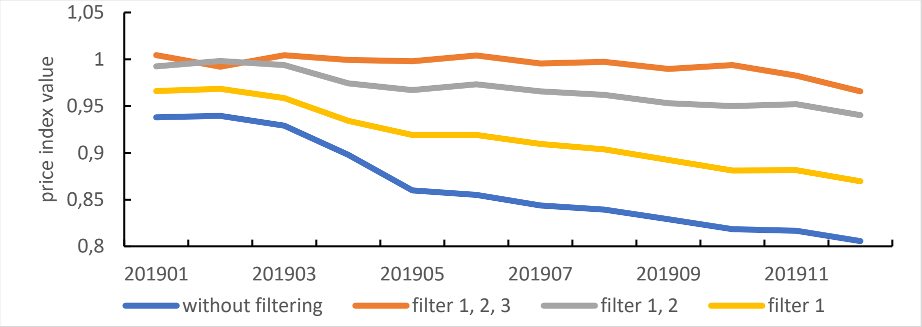

Fig. 3.Impact of the outlier filter (filter 1), dumping filter (filter 2) and low sale filter (filter 3) on the Chain Jevons index value based on example of products of the homogeneous ECOICOP6 group - biscuits and wafers without filling, Dec. 2018 – Dec. 2019

As previously stated, the scanner data allows us to compile unweighted and weighted indices at the elementary level, and the selection of formula and method can have a significant impact on the results obtained. Differences between individual indices considered are shown in Fig. 2 and 3. It is no surprise that the values of the Carli index are above the values of the Dutot index and the Jevons index. Which also follows from their mathematical properties. [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]

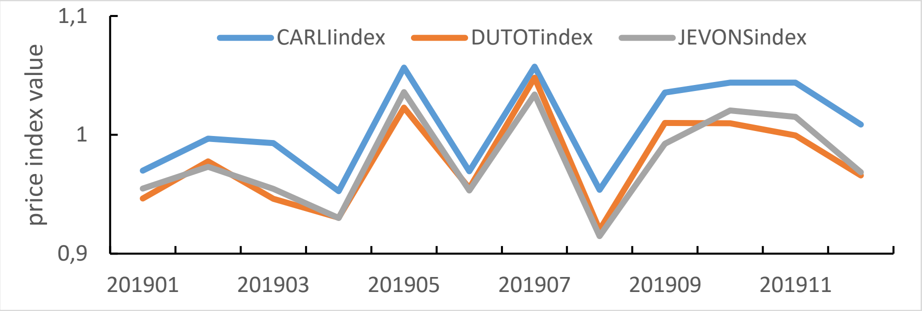

Fig.2 Comparison of the development of unweighted bilateral chained indices based on the example of products of the homogeneous group ECOICOP6 - fresh butter, Dec. 2018 - Dec. 2019

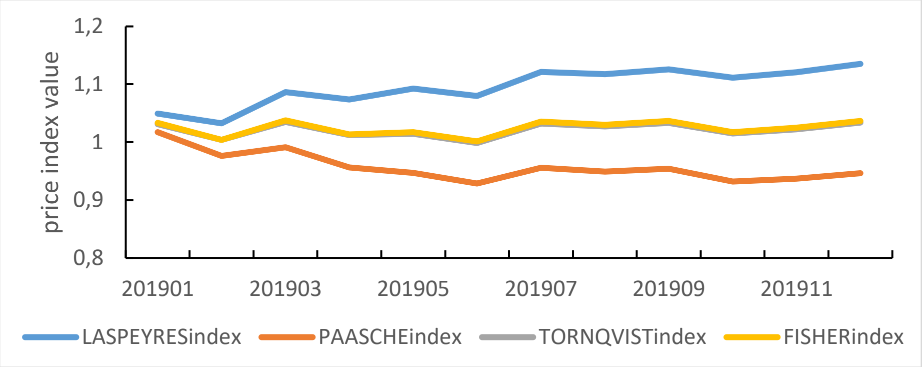

Fig. 3. Comparison of the development of weighted bilateral chained indices based on the example of products of the homogeneous group ECOICOP6 - fresh whole milk, Dec. 2018 - Dec. 2019

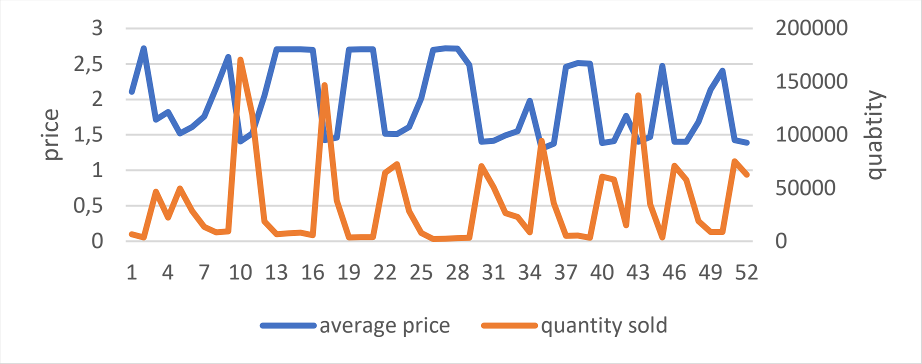

Fig. 4. Comparison of the development of the price and the quantity of product sold- butter traditional Koliba, sold in one of the supermarket chains

Superlative indices are based on economic theory and the weights used reflect the changes in consumer behaviour illustrated in Fig. 3. In the fixed basket approach, superlative indices (Fischer, Tornqvist) approximate each other [**Chyba! Nenašiel sa žiaden zdroj odkazov.**]. In our case (dynamic basket, chained versions of indices) they behave in the same way. Their development is in line with the economic approach in the price index theory, according to which superlative indices approximate the cost of living index and Laspeyres index provides an upper limit and Paasche a lower limit to the COLI.

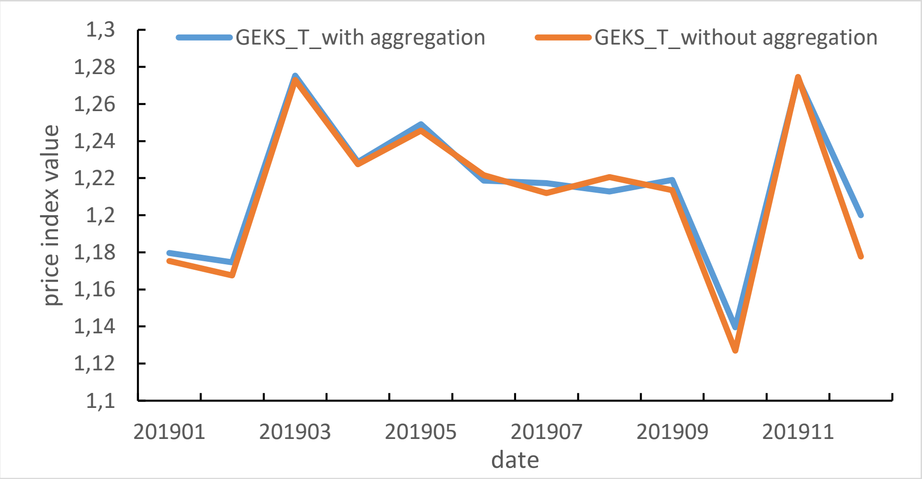

Fig. 5. Impact of the aggregation over supermarket chains on the multilateral GEKS- Törnqvist index results calculated (ECOICOP6 – rice husked, Dec. 2018 – Dec. 2019).

One of the questions that needs to be answer in relation to the implementation of scanner data into CPI production is the question, whether the indices at elementary level will be compiled above common database of products or above databases of individual retail chains and then aggregate them (Laspeyres formula). In the conditions of the Slovak Republic, the results of the experimental study in this area confirmed insignificant differences between the mentioned procedures in the division of food and non-alcoholic beverages (see Fig. 5), and therefore, a simpler procedure is recommended, and that is to implement the compilation of CPI over a common product database.

4 Conclusion

The Slovak Republic is still in the process of preparation for the usage of scanner data in the official statistics. Some issues have remained unresolved, and they will require further research and experimentation with real transaction data.

In addition, IT system which would combine and analyze data from various sources is under preparation. Despite of this, several experiments on data processing, classification, matching (against the base period or the previous period) and calculating different price indices on real data using SAS or R package software have been performed.

Based on the performed experiments, the same homogeneous product groups for all cooperating retail chains in the division of food and non-alcoholic beverages were made up. Indices at elementary level will be calculated over a common database of matched products, i.e. without aggregation over supermarket chains. As a result of the ongoing experimental study, the threshold values have been set to exclude extreme price changes and clearance goods. It has been stopped dealing with the compilation of CPI over a fixed consumer basket (selection of products from scanner data). Exclusively the dynamic approach will be the subject of the other experiments. However, the issue of implementing scanner data into CPI statistical production is very broad and multifaceted issue and many methodological questions are still unanswered. For example, choosing the right index formula remains the great challenge.

Ongoing experiments will be focused mainly on the possible implementation of the Jevos index, superlative chained indices and especially multilateral indices

References

1. Bialek, J., Roszko-Wójtowicz, E.: The Impact of the Price Index Formula on the Consumer Price Index Measurement. Statistica: Statistics and economy journal 99 (3), 246–258 (2019).
2. Bialek, J.: PriceIndices – A New R Package for Bilateral and Multilateral Price Index Calculations. Statistica: Statistics and economy journal 101 (2), 54–69 (2020).
3. Bialek, J.: Remarks on Price Index Methods for the CPI Measurement Using Scanner data. Statistica: Statistics and economy journal 100 (1), 122–141 (2021)
4. CIRCABC Homepage, https://circabc.europa.eu/sd/a/8e1333df-ca16-40fc-bc6a-1ce1be37247c/Practical-Guide-Supermarket-Scanner-Data-September-2017.pdf, last accessed 2019/05/13.

5. Diewert, W.E.: Exact and Superlative Index Numbers. Journal of Econometrics 4, 114-145 (1976).
6. Chessa, A.: A new methodology for processing scanner data in the Dutch CPI. In.: Eurostat review of National Accounts and Macroeconomic Indicators, 1, 49-69 (2016).
7. International Labour Organization Homepage, http://www.ilo.org/public/english/bureau/stat/download/cpi/corrections/annex1.pdf, last accessed 2021/08/12
8. Ivancic, L., Diewert, W. E., Fox, K. J.: Scanner Data, Time Aggregation and the Construction of Price Indices. In: Journal of Econometrics, 161(1), 24–35 (2011).
9. National Bureau of Economic Research Homepage, http://www.nber.org/papers/w27144, last accessed 2021/08/12.
10. UNECE Homepage, https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2014/WS1/WS1_1_Diewert_on_Diewert_Consumer_Price_Statistics__in_the_UK_v.7__06.08__Final.pdf, last accessed 2021/08/12.
11. Van Loon, K. V., Roels, D.: Integrating big data in the Belgian CPI. In: The meeting of the group of experts on consumer price indices, Geneva, Switzerland, (8–9 May, 2018).